

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-320129

(43) 公開日 平成10年(1998)12月4日

(51) Int.Cl.<sup>6</sup>

G 0 6 F 3/06

識別記号

3 0 4

5 4 0

F I

G 0 6 F 3/06

3 0 4 B

5 4 0

審査請求 未請求 請求項の数 3 O L (全 9 頁)

(21) 出願番号

特願平9-129725

(22) 出願日

平成9年(1997)5月20日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 平野 正人

神奈川県小田原市国府津2880番地 株式会社

日立製作所ストレージシステム事業部内

(74) 代理人 弁理士 筒井 大和

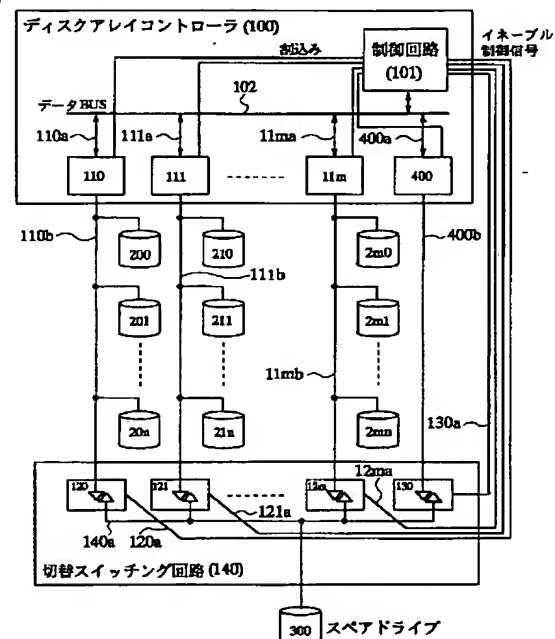
(54) 【発明の名称】 ディスクアレイ装置

(57) 【要約】

【課題】 制御系の多重障害によるデータ喪失、およびスペアドライブの接続による性能低下等を防止する。

【解決手段】 ディスクアレイコントローラ100内に設けられた複数のディスクコントローラ110～11mの各々の下位バス110b～11mbに、複数のディスクドライブ200～20n、ディスクドライブ2m0～2mnを系列的に接続し、スペアコントローラ400およびスペアドライブ300と、下位バス110b～11mbおよびスペアコントローラ400の下位バス400bに対応した入出力バッファ120～12m、130を備えた切替スイッチング回路140を設け、ディスクコントローラ110～11mやディスクドライブ200～2mnの障害時に対応する入出力バッファ120～12m、130を選択的にイネーブルにして、スペアコントローラ400やスペアドライブ300を障害系列に接続する。

図 1



(2)

特開平10-320129

**【特許請求の範囲】**

**【請求項1】** 複数のディスクコントローラと、前記ディスクコントローラの各々に個別に接続される複数のバスと、複数の前記バスのいずれかを介して前記ディスクコントローラに接続される複数のディスクドライブと、少なくとも一つのスペアドライブと、任意の契機にて前記スペアドライブを複数の前記バスの任意の一つに選択的に接続するスイッチ手段と、を含むことを特徴とするディスクアレイ装置。

**【請求項2】** 複数のディスクコントローラと、前記ディスクコントローラの各々に個別に接続される複数のバスと、複数の前記バスのいずれかを介して前記ディスクコントローラに接続される複数のディスクドライブと、少なくとも一つのスペアコントローラと、任意の契機にて前記スペアコントローラを複数の前記バスの任意の一つに選択的に接続するスイッチ手段と、を含むことを特徴とするディスクアレイ装置。

**【請求項3】** 複数のディスクコントローラと、前記ディスクコントローラの各々に個別に接続される複数のバスと、複数の前記バスのいずれかを介して前記ディスクコントローラに接続される複数のディスクドライブと、少なくとも一つのスペアドライブと、少なくとも一つのスペアコントローラと、任意の契機にて前記スペアドライブおよび前記スペアコントローラの各々を複数の前記バスの任意の一つに選択的に接続するスイッチ手段と、を含むことを特徴とするディスクアレイ装置。

**【発明の詳細な説明】****【0001】**

**【発明の属する技術分野】** 本発明は、ディスクアレイ技術に関し、特に、複数のディスクコントローラによって複数のディスクドライブを制御する構成にてRAID (Redundant Array of Inexpensive Disks) システムを実現するディスクアレイ装置等に適用して有効な技術に関する。

**【0002】**

**【従来の技術】** たとえば、情報処理システムの分野では、外部記憶装置として、冗長構成の比較的安価な小形ディスクドライブ群にて一つの記憶システムを構築するRAID (ディスクアレイ) 技術が知られている。このディスクアレイ技術では、書込データを幾つかに分割し、さらに分割された書込データから冗長データを生成し、これらの分割された書込データおよび冗長データのグループを個別のディスクドライブに分散して並列転送することにより、見かけ上の入出力速度の向上を図るとともに、データ障害時には、健全な残りのデータと冗長データとから障害データの復元を行うことでデータの信頼性の確保を実現している。

**【0003】** このようなディスクアレイ技術については、たとえば、特開平7-114445号公報に開示された技術が知られている。この技術では、複数のディス

クアダプタ (ディスクコントローラ) の各々に独立に設けられた複数のポート (バス) の各々に複数のディスクユニット (ドライブ) を系列的に接続し、分割された複数のデータブロックおよび当該データブロック群から生成された冗長データとしてのパリティを各ディスクアダプタに属するディスクユニットの系列を横断する方向に分散して格納するディスクアレイの構成が示されている。また、ホスト計算機から与えられる論理アドレスから、ディスクアレイ装置におけるディスクドライブ群の構成を意識した物理アドレスへの変換の一手法が述べられている。

**【0004】**

**【発明が解決しようとする課題】** 上述の従来のディスクアレイでは、下記のような技術的課題があった。

**【0005】** 第1に、ディスクコントローラに障害が発生した場合は、そのディスクコントローラのポートに接続されているディスクドライブが使用不可能な状態、すなわちディスクドライブの冗長度がない縮退状態になる。縮退状態になると読み出し時には障害ディスクドライブのデータを他のディスクドライブのデータから生成させるため、ドライブアクセスの処理時間が大幅に増大してしまう。

**【0006】** 第2に、ディスクコントローラ障害の状態では次のディスクコントローラに障害が発生した場合にはデータが失われてしまう。

**【0007】** 第3に、個々のディスクドライブの障害に備えて、たとえばスペアドライブを1つのディスクコントローラに固定的に接続した場合、ディスクドライブに障害が発生すると、縮退状態を回避するため障害ディスクドライブのデータをスペアドライブに復旧させるが、スペアドライブが接続されているディスクコントローラの系列以外のディスクドライブに障害が発生した場合、スペアドライブが接続されたディスクコントローラはディスクドライブが1台多く接続された状態となり、当該ディスクコントローラに余分な負荷がかかってしまい、データ復旧の処理時間が増大してしまう。

**【0008】** 第4に、上述の第3の項におけるデータ復旧の処理後の通常の稼働時においても、スペアドライブに接続されたディスクコントローラはディスクドライブが1台多く接続された状態となり、コントローラに負荷がかかってしまい、ドライブアクセスの処理時間が顕著に増大してしまう。

**【0009】** 本発明の目的は、ディスクコントローラの障害に起因する縮退運転の発生を回避して、縮退運転に起因するデータ転送速度の低下を防止することが可能なディスクアレイ技術を提供することにある。

**【0010】** 本発明の他の目的は、ディスクコントローラの多重障害に起因するデータ喪失を防止することが可能なディスクアレイ技術を提供することにある。

**【0011】** 本発明の他の目的は、スペアドライブの接

(3)

特開平10-320129

続に起因する特定のディスクコントローラへの負荷の偏りを回避してディスクドライブの障害に起因するデータ復旧の所要時間を短縮することが可能なディスクアレイ技術を提供することにある。

【0012】本発明の他の目的は、スベアドライブの接続に起因する特定のディスクコントローラへの負荷の偏りを回避して、稼働時のデータ転送速度を向上させることが可能なディスクアレイ技術を提供することにある。

【0013】

【課題を解決するための手段】本発明は、複数のディスクコントローラと、ディスクコントローラの各々に個別に接続される複数のバスと、複数のバスのいずれかを介してディスクコントローラに接続される複数のディスクドライブとを含む構成のディスクアレイにおいて、少なくとも一つのスベアドライブと、このスベアドライブを障害発生時等の任意の契機にて、複数のバスの任意の一つに選択的に接続するスイッチ手段を備えたものである。

【0014】また、本発明は、複数のディスクコントローラと、前記ディスクコントローラの各々に個別に接続される複数のバスと、複数の前記バスのいずれかを介して前記ディスクコントローラに接続される複数のディスクドライブとを含む構成のディスクアレイにおいて、少なくとも一つのスベアコントローラと、このスベアコントローラを任意の契機にて複数の前記バスの任意の一つに選択的に接続するスイッチ手段を備えるようにしたものである。

【0015】また、本発明は、複数のディスクコントローラと、ディスクコントローラの各々に個別に接続される複数のバスと、複数のバスのいずれかを介してディスクコントローラに接続される複数のディスクドライブとを含む構成のディスクアレイにおいて、少なくとも一つのスベアドライブ、および少なくとも一つのスベアコントローラと、このスベアドライブおよびスベアコントローラの各々を任意の契機にて複数のバスの任意の一つに選択的に接続するスイッチ手段を備えるようにしたものである。

【0016】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照しながら詳細に説明する。

【0017】図1は、本発明の実施の形態であるディスクアレイ装置の構成の一例を示す概念図である。

【0018】本実施の形態のディスクアレイ装置は、たとえば、共通のデータバス102に上位バス110a～111maを介して接続された複数のディスクコントローラ110～111mを含むディスクアレイコントローラ100と、複数のディスクコントローラ110～111mの各々に接続される下位バス110b～111mbの各々に系列的に接続されている複数のディスクドライブ200～20n、ディスクドライブ210～21n、・・・デ

ィスクドライブ2m0～2mnと、を含む構成となっている。

【0019】複数のディスクドライブ200～2mnとディスクアレイコントローラ100とが、たとえばSCSI規格等のインターフェイスにて接続される場合、ディスクコントローラ110～111mはSCSIコントローラにて構成され、下位バス110b～111mbはSCSIバスで構成される。

【0020】本実施の形態の場合、ディスクアレイコントローラ100は、たとえば図示しない上位装置から受領したデータを、系列数m個に分割するとともに当該分割データ群からパリティ等の冗長データを生成してパリティグループを構成し、このパリティグループを構成するm+1個のデータを、ディスクコントローラ110～111mの配下のm+1台の、たとえばディスクドライブ200、210、・・・2m0に並列に転送して格納する動作を行う。また、パリティグループ内の分割データのリード時にエラーが発生した場合には、当該パリティグループ内の他の分割データと冗長データとから障害の分割データの復元処理を行う。

【0021】本実施の形態の場合、ディスクアレイコントローラ100には、上位バス400aを介して、他のディスクコントローラ110～111mと等価にデータバス102に接続されるスベアコントローラ400が設けられている。

【0022】さらに、本実施の形態の場合には、複数のディスクドライブ200～2mnの他にスベアドライブ300が設けられており、このスベアドライブ300は、切替スイッチング回路140を介して、ディスクコントローラ110～111mの下位バス110b～111mbのいずれか一つに、任意の契機にて選択的に接続可能な構成となっている。同様に、スベアコントローラ400の下位バス400bは、ディスクコントローラ110～111mの下位バス110b～111mbのいずれか一つに、任意の契機にて接続可能になっている。

【0023】すなわち、切替スイッチング回路140では各ディスクコントローラ110～111m、およびスベアコントローラ400の接続部分に入出力バッファ120～12m、および入出力バッファ130が設けられており、これらの入出力バッファ120～12m、および入出力バッファ130は、スイッチング回路内部バス140aを介して相互に共通に接続されているとともに、スベアドライブ300に対しても共通に接続されている。

【0024】個々の入出力バッファ120～12m、および入出力バッファ130は、制御線120a～12maおよび制御線130aによってイネーブル/ディセーブルが制御され、イネーブルの状態では、対応する下位バス110b～111mbおよび下位バス400bと、スイッチング回路内部バス140aとが接続され、ディセ

(4)

特開平10-320129

ケーブルでは遮断される。

【0025】すなわち、入出力バッファ120～12mのどれか一つをイネーブルにすることにより、対応する下位バス110b～11mb（ディスクコントローラ110～11m）の一つが選択的にスベアドライブ300に接続される状態となる。

【0026】また、入出力バッファ130と、入出力バッファ120～12mのどれか一つを同時にイネーブルにすることによって、スベアコントローラ400と、下位バス110b～11mbのいずれか一つとが接続された状態となる。

【0027】本実施の形態の場合、制御線120a～12maおよび制御線130aは、ディスクアレイコントローラ100の内部に設けられた制御回路101に接続されており、この制御回路101から入出力バッファ120～12mおよび入出力バッファ130に与えられるイネーブル制御信号によって個々の入出力バッファ120～12mおよび入出力バッファ130のイネーブル／ディセーブルが制御される構成となっている。

【0028】すなわち、各ディスクコントローラ110～11m、およびスベアコントローラ400は制御回路101へ割込み信号を送信し、制御回路101からのイネーブル制御信号の選択的な発行を促すことにより、切替スイッチング回路140における下位バス110b～11mbおよび下位バス400bの各々の接続の有無の制御を行う。

【0029】以下、本実施の形態のディスクアレイ装置における障害発生時の作用の一例について説明する。

【0030】①任意のディスクドライブ（たとえばディスクドライブ210）に障害が発生した場合の本実施の形態のディスクアレイ装置の状態の一例を図2に、その時の動作のフローチャートの一例を図6に示す。ディスクコントローラ111は、その下位バス111bに接続されている系列下のディスクドライブ210から正常な応答がない（例えばデータエラー、タイムアウト等）場合は、ディスクアレイコントローラ100内の制御回路101に割込み信号を送る。

【0031】この割込みを受け、ディスクアレイコントローラ100はディスクコントローラ111のステータスを読み取り、ディスクドライブ障害であることを検出する。ディスクドライブ障害を検出すると、制御回路101は障害のディスクドライブ210が接続されている下位バス111bの入出力バッファ121のみを選択的にイネーブルにしてスベアドライブ300を障害のディスクドライブと同じ系列の下位バス111bに接続させる。スベアドライブ300には、同じパリティグループ内の他の正常なディスクドライブ200、220～2mのデータから障害のディスクドライブ210のデータを復元してコピー（データ復旧）を行うが、スベアドライブ300には障害のディスクドライブ210の接続さ

れていたディスクコントローラ111が接続されるため、各ディスクコントローラ110～11mは同じパリティグループ内ではそれぞれ1台のディスクドライブが接続されることになり、たとえば特定の系列にスベアドライブを固定的に接続する従来の場合等に比較して、特定のディスクコントローラに負荷が偏ることはない。従って、スベアドライブ300に対する迅速な障害データの復旧処理を行うことが可能になるとともに、データ復旧後の通常の稼働時においても、障害に前後におけるデータの並列転送の状態に変化はなく、スベアドライブ300の接続に起因する性能低下の発生もない。

【0032】なお、複数のスベアドライブ300～302を接続させた場合の構成の一例を図3に、障害時の動作のフローチャートを図9に例示する。この場合、複数のスベアドライブ300～302の各々は、入出力バッファ150～152を介して、スイッチング回路内部バス140aに接続されている。この入出力バッファ150～152は、制御線150a～152aを介して制御回路101にてイネーブル／ディセーブルが制御される。

【0033】この図3の構成例では、上記手順によるデータ復旧中にスベアドライブ300に障害が発生した場合でも、図9のフローチャートに例示されるように、現在のスベアドライブ300のバス（入出力バッファ150）を切り離し、別のスベアドライブ301～302のバス（入出力バッファ151～152）をイネーブルにすることでドライブ閉塞状態になることを防止でき、ディスクアレイ装置の信頼性がより向上する。

【0034】②任意のディスクコントローラ（たとえばディスクコントローラ11m）に障害が発生した場合、本実施の形態のディスクアレイ装置の状態の一例を図4に、その時の動作の一例のフローチャートを図7に示す。ディスクアレイコントローラ100はそのディスクコントローラ11mから正常な応答がない場合にディスクコントローラ障害であることを検出する。ディスクコントローラ障害を検出すると、制御回路101はスベアコントローラ400が接続されている下位バス400bに対応した入出力バッファ130、および障害の発生したディスクコントローラ11mが接続されている下位バス11mbの入出力バッファ12mをイネーブルにして、スベアコントローラ400を障害のディスクコントローラ11mと同じ下位バス11mbに接続させる。そしてディスクアレイコントローラ100は、障害のディスクコントローラ11mにおける制御情報等の設定をスベアコントローラ400に反映させる。スベアコントローラ400は接続された下位バス11mbの系列のディスクドライブ2m0～2mnを制御する。この場合、スベアドライブ300へのアクセスはスベアコントローラ400のみになる。

【0035】③ディスクアレイコントローラ100の内

(5)

特開平10-320129

部で、データバス102とディスクコントローラ110～11mとの間の上位バス110a～11maに障害（断線、接触不良等）が発生した場合の本実施の形態のディスクアレイ装置の状態の一例を図5に、また、制御動作の一例を図8のフローチャートに示す。この図5に例では、ディスクコントローラ110の上位バス110aに障害が発生した場合が例示されている。ディスクアレイコントローラ100はデータバス102のパリティチェック等によりバス障害であることを検出する。この時、制御回路101はスペアコントローラ400が接続されている下位バス400bの入出力バッファ130、および障害が発生した上位バス110a（ディスクコントローラ110）に対応した下位バス110bの入出力バッファ120をイネーブルにしてスペアコントローラ400を障害の発生した上位バス110aと同系列の下位バス110bと接続させる。

【0036】これにより、下位バス110bの系列のディスクドライブ200～20nは、スペアコントローラ400によって正常に制御される。この時、ディスクコントローラ110が持つ構成情報等をスペアコントローラ400に与えることは前記②の場合と同様である。また②の場合と同様に、スペアドライブ300へのアクセスはスペアコントローラ400のみになる。

【0037】このように、本実施の形態のディスクアレイ装置においては、②のディスクコントローラ110～11mのいずれかに障害が発生した場合や、③の上位バス110a～11maのいずれかに障害が発生した場合には、スペアコントローラ400に切り換えることにより、縮退状態に陥ることがなく可用性や信頼性が向上するとともに、データ転送速度等の性能のそのまま維持できる、という利点がある。また、スペアコントローラ400の使用中に、さらにディスクコントローラ110～11mや上位バス110a～11ma等に障害が発生しても縮退状態に移行するだけであり、データの喪失は発生せず、信頼性の維持向上を実現することができる。

【0038】なお、特に図示しないが、スペアコントローラ400を複数設けることも本発明に含まれる。この場合には、対応する下位バスおよび入出力バッファを増やせばよく、上述の②と同様の制御にて耐故障性能がより向上する。

【0039】図10は、本実施の形態のディスクアレイ装置における上述の各障害時の動作を含む作用の一例を示すフローチャートである。

【0040】すなわち、データ入出力処理（ステップ501）を、エラーの有無を監視しつつ（ステップ502）継続し、エラー検出の場合には、要因を切りわけ（ステップ503、ステップ504、ステップ505）、ドライブ起因の場合には、空きのスペアドライブの有無を判別し（ステップ507）、有りの場合には、上述の図6または9の処理を実行し（ステップ50

8）、無い場合には縮退処理を実行して（ステップ509）、ステップ501に戻って稼働を継続する。

【0041】同様に、ディスクコントローラや、ディスクアレイコントローラ100内部のバスの障害の場合には、空きのスペアコントローラの有無を判別し（ステップ510）、空きのスペアコントローラがある場合には前述の図7または図8の処理を実行し（ステップ511）、無い場合には縮退処理を実行して（ステップ512）、ステップ501に戻って稼働を継続する。

【0042】上記以外の障害の場合には、対応する所定の処理を行い（ステップ506）、ステップ501に戻って稼働を継続する。

【0043】以上の動作により、本実施の形態のディスクアレイ装置によれば、上述のようなスペアドライブの接続における性能低下の防止やコントローラ系の多重障害における性能低下およびデータ喪失の防止等の優れた効果を得ることができる。

【0044】以上本発明者によってなされた発明を実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【0045】たとえば、ディスクコントローラとディスクドライブとの接続インターフェイスとしては、上述の実施の形態に例示したSCSIに限らず、一般のインターフェイスを使用することができる。

【0046】

【発明の効果】本発明のディスクアレイ装置によれば、ディスクコントローラの障害に起因する縮退運転の発生を回避して、縮退運転に起因するデータ転送速度の低下を防止することができる、という効果が得られる。

【0047】また、本発明のディスクアレイ装置によれば、ディスクコントローラの多重障害に起因するデータ喪失を防止することができる、という効果が得られる。

【0048】また、本発明のディスクアレイ装置によれば、スペアドライブの接続に起因する特定のディスクコントローラへの負荷の偏りを回避してディスクドライブの障害に起因するデータ復旧の所要時間を短縮することができる、という効果が得られる。

【0049】また、本発明のディスクアレイ装置によれば、スペアドライブの接続に起因する特定のディスクコントローラへの負荷の偏りを回避して、稼働時のデータ転送速度を向上させることができる、という効果が得られる。

【図面の簡単な説明】

【図1】本発明の実施の形態であるディスクアレイ装置の構成の一例を示す概念図である。

【図2】本発明の実施の形態であるディスクアレイ装置におけるドライブ障害時の状態の一例を示す概念図である。

【図3】本発明の実施の形態であるディスクアレイ装置

(6)

特開平10-320129

において複数のスペアドライブを設けた場合のドライブ障害時の状態の一例を示す概念図である。

【図4】本発明の実施の形態であるディスクアレイ装置におけるディスクコントローラ障害時の状態の一例を示す概念図である。

【図5】本発明の実施の形態であるディスクアレイ装置におけるバス障害時の状態の一例を示す概念図である。

【図6】本発明の実施の形態であるディスクアレイ装置におけるドライブ障害時の作用の一例を示すフローチャートである。

【図7】本発明の実施の形態であるディスクアレイ装置におけるディスクコントローラ障害時の作用の一例を示すフローチャートである。

【図8】本発明の実施の形態であるディスクアレイ装置におけるバス障害時の作用の一例を示すフローチャートである。

【図9】本発明の実施の形態であるディスクアレイ装置

において複数のスペアドライブを設けた場合のドライブ障害時の作用の一例を示すフローチャートである。

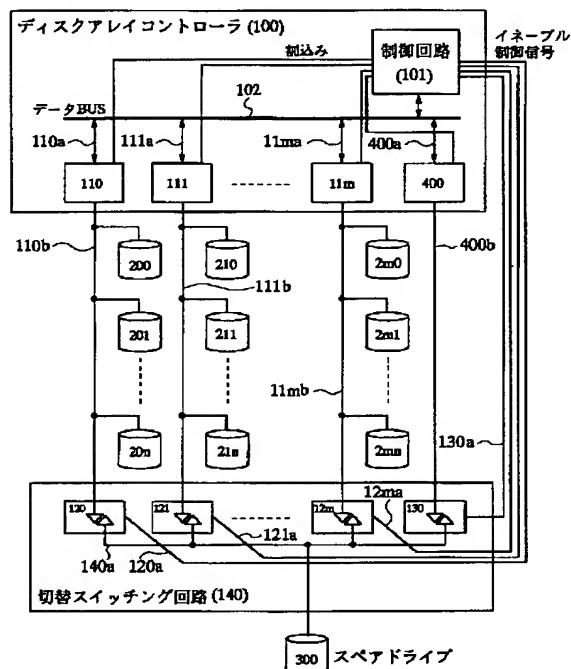
【図10】本発明の一実施の形態であるディスクアレイ装置における各種障害時の動作を含む作用の一例を示すフローチャートである。

#### 【符号の説明】

100…ディスクアレイコントローラ、101…制御回路、102…データバス、110～11m…ディスクコントローラ、110a～11ma…上位バス、110b～11mb…下位バス、120～12m…入出力バッファ、120a～12ma…制御線、130…入出力バッファ、130a…制御線、140…切替スイッチング回路（スイッチ手段）、140a…スイッチング回路内部バス、150～152…入出力バッファ、150a～152a…制御線、200～2mn…ディスクドライブ、300～302…スペアドライブ、400…スペアコントローラ、400a…上位バス、400b…下位バス。

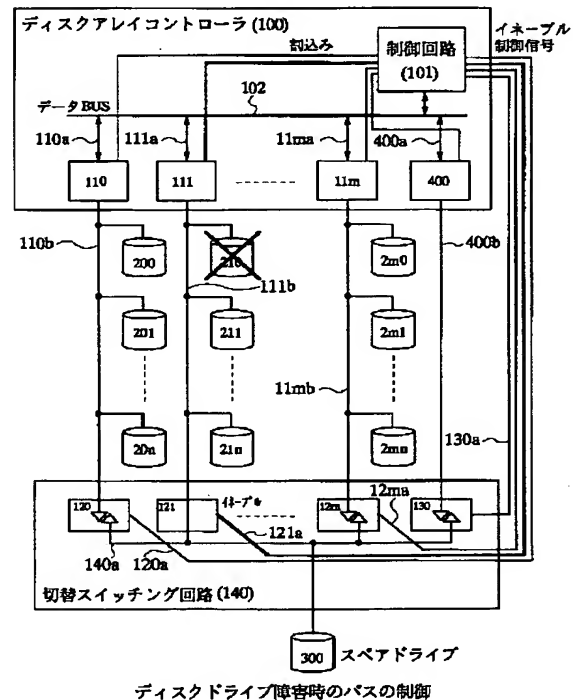
【図1】

図 1



【図2】

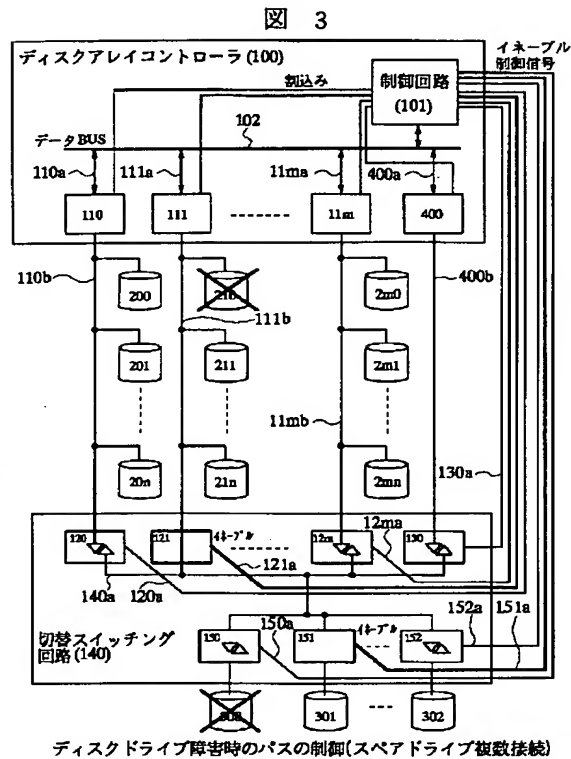
図 2



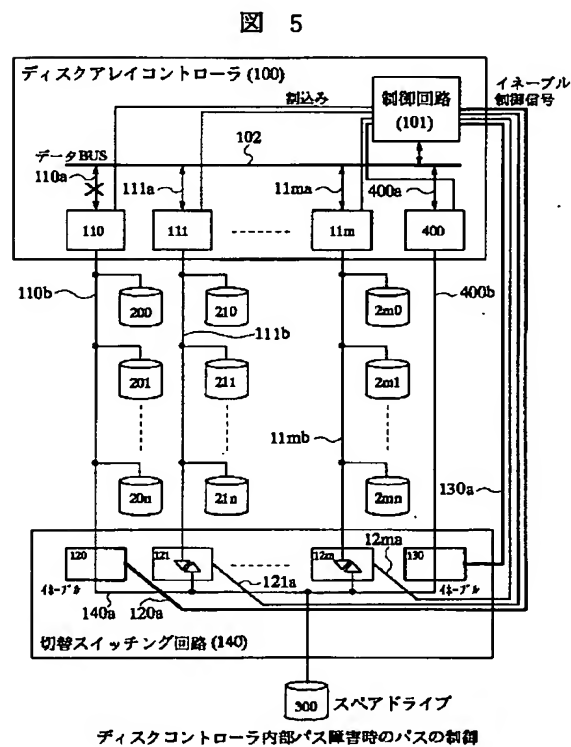
(7)

特開平10-320129

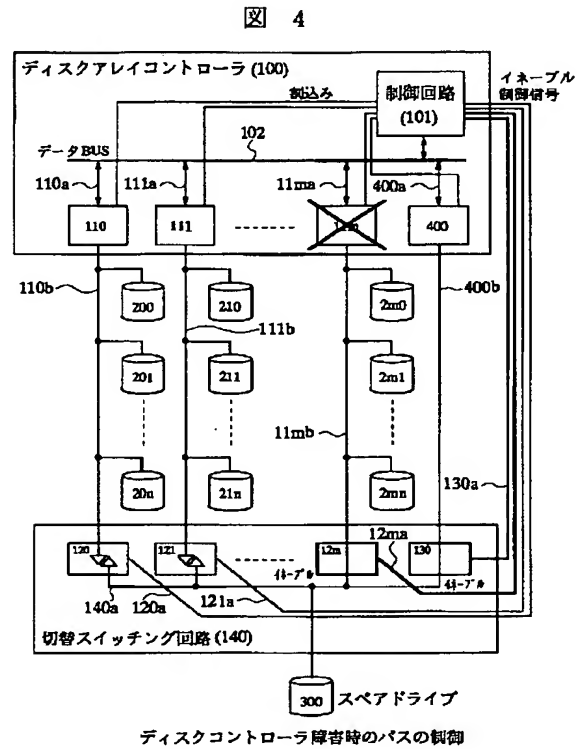
【図3】



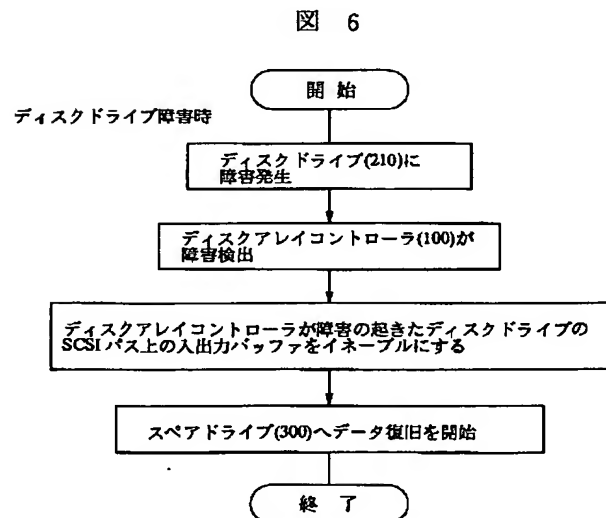
【図5】



【図4】



【図6】

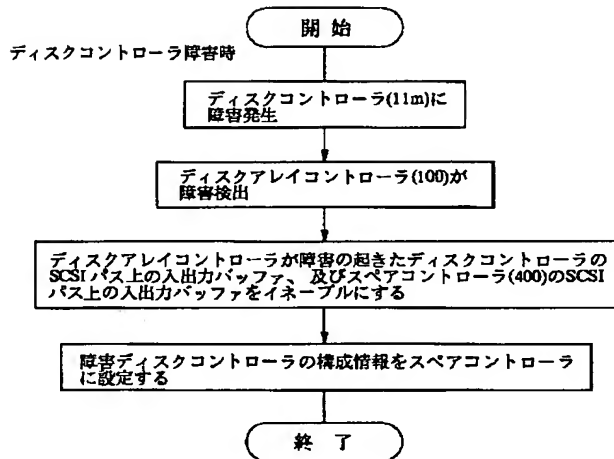


(8)

特開平10-320129

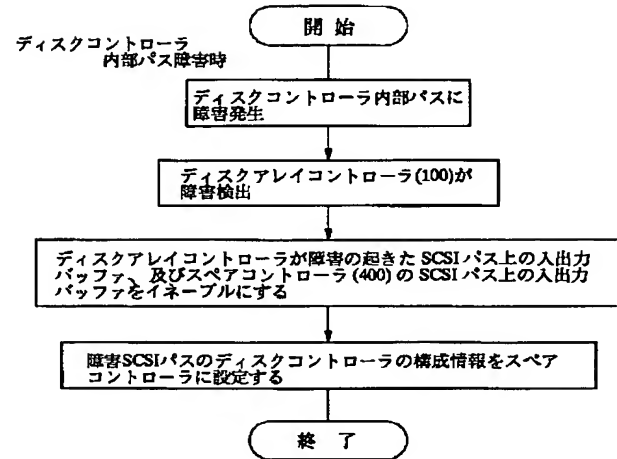
【図7】

図 7



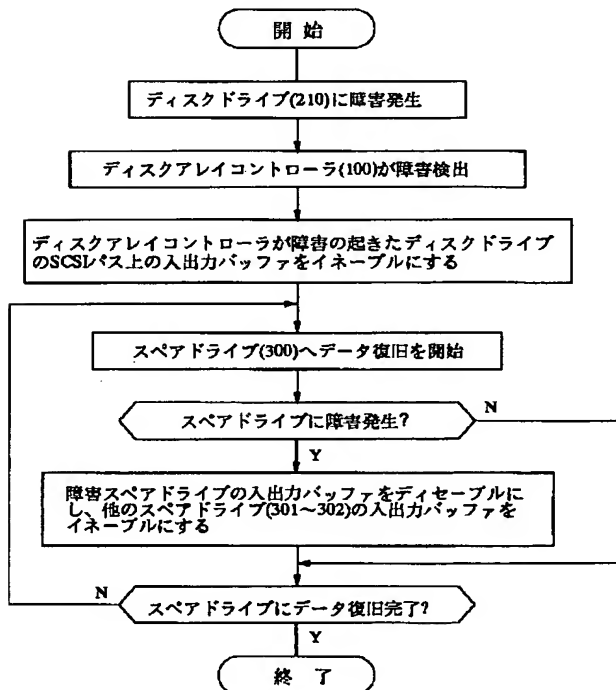
【図8】

図 8



【図9】

図 9





(9)

特開平10-320129

【図10】

図 10

